TAMPERE ECONOMIC WORKING PAPERS

REVERSAL OF THE KUZNETS CURVE: STUDY ON THE INEQUALITY-DEVELOPMENT RELATION USING TOP INCOME SHARES DATA

Elina Tuominen

Working Paper 110
September 2016

SCHOOL OF MANAGEMENT

FI-33014 UNIVERSITY OF TAMPERE, FINLAND

ISSN 1458-1191
ISBN 978-952-03-0252-8

Reversal of the Kuznets curve: Study on the inequality–development relation using top income shares data[☆]

Elina Tuominen

University of Tampere, Finland (elina.tuominen@uta.fi)

September 2016

Abstract

In this study, recently published top 1% income share series are exploited in studying the inequality–development association in 26 countries from 1900 to 2010. The top income shares data are of high quality and provide interesting possibilities for studying slow development processes. Because many empirical inequality–development studies have challenged the use of quadratic specifications, this study addresses the issue of functional form by applying penalized spline methods. The relationship between the top 1% income share and development is found to experience a reversal at the highest levels of development and, thus, a positive association is now observed in many "advanced" economies. In an additional analysis covering a shorter time period, the discovered positive relationship holds at the highest levels of development when controls for two sectoral measures are included.

Keywords: inequality, top incomes, development, nonlinearity, longitudinal data
JEL classification: N30, O11, O15

Acknowledgments

This study has received funding from the Finnish Cultural Foundation and the Academy of Finland (project 268863). The author thanks Markus Jäntti, Ravi Kanbur, Arto Luoma, and Matti Tuomala for the discussions. The study has also benefited from comments at the Finnish Doctoral Programme in Economics (FDPE) Workshops, the Annual Meeting of the Finnish Economic Association 2014, the Annual Meeting of the Austrian Economic Association 2014, the IIPF Congress 2014, and the UNU-WIDER Conference 2014 on "Inequality—measurement, trends, impacts, and policies." Remaining errors are the author's own.

[☆] An earlier version of this study was published by UNU-WIDER, Helsinki (UNU-WIDER Working Paper 2015/036).

1. Introduction

In his seminal paper, Kuznets (1955) presented the famous "inverted-U hypothesis" between inequality and economic development; inequality first increases and then decreases as the country develops.[1] He suggested that during this process, the focus of the economy shifts from agriculture to modern sectors.[2] In addition to this famous idea of a sectoral shift, Kuznets discussed various other factors that affect the income distribution during the development process. For example, he noted that the concentration of savings at the top of the distribution induces inequality in the distribution before taxes and transfers, and he discussed equalizing forces such as political pressure for redistribution. Subsequently, various theoretical models have generated a Kuznets-type curve (e.g., Robinson, 1976; Greenwood & Jovanovic, 1990; Galor & Tsiddon, 1996; Aghion & Bolton, 1997; Dahan & Tsiddon, 1998). Empirical studies have presented mixed evidence on the shape of the inequality–development association, and the debate has focused on whether the results support the inverse-U hypothesis. A short and selective introduction to the empirical literature is provided next.[3]

In empirical applications, the chosen functional form plays an important role. For example, a cross-sectional study by Ahluwalia (1976) supports the inverted-U link, but Anand and Kanbur (1993) challenge the data quality and chosen functional forms. In comparison, Huang (2004), Lin et al. (2006), and Huang and Lin (2007) apply nonparametric methods to cross-sectional data and find evidence for the Kuznets hypothesis. However, it is possible that cross-sectional data cannot capture the complexity of the process. Panel studies have become more common after the construction of new inequality data sets. Possibly the most famous panel data set is by Deininger and Squire (1996). Although these data have been exploited in several studies, parametric analyses have shown differing results (e.g., Deininger & Squire, 1998; Barro, 2000). Further, Atkinson and Brandolini (2001) demonstrate that also this inequality data set has its shortcomings.

[1] Using data from the United States, Germany, and the United Kingdom, Kuznets (1955) got an impression of constancy in inequality around the turn of the twentieth century, followed by a secular decline in inequality at least since the 1920s.

[2] Kuznets (1955) provided numerical illustrations where (under certain assumptions) a mere population shift from the rural to urban sector can affect the overall income distribution: inequality first increases, and then declines.

[3] Further, Fields (2001) and Frazer (2006) provide overviews of the literature.

Recent studies suggest that using flexible methods is well-founded in inequality–development investigations. Frazer (2006) applies nonparametric regression in his study that spans approximately 50 years. In his pooled models, he discovers a nonlinear Gini–development association that is more complex than a second-degree polynomial. Specifically, he finds that the curve may be flat before it experiences a negative slope. His illustrations also show that the association may turn positive at the highest levels of development, but the confidence interval becomes wide at these development levels. Moreover, Zhou and Li (2011) conduct a nonparametric investigation on the inequality–development association using unbalanced panel data for the period 1962–2003. They find an inverse-U relation between Gini coefficients and economic development, but only after a certain level of development is reached. Further, Desbordes and Verardi (2012) use semiparametric methods with Gini data for the 1960–2000 period and provide empirical evidence for the latter stages of the Kuznets-type relation. Desbordes and Verardi also show that misspecified functional forms can lead to differing results on the inequality–development association.

Various inequality indices—including top income shares—have shown an upward trend in many countries over the last 20–30 years, and the inverse-U association has been challenged. In addition, List and Gallet (1999) study an unbalanced panel from 1961 to 1992 and find that, at the highest levels of economic development, there is a positive correlation between inequality and development. Although List and Gallet admit that the positive association may be a result of various factors, they suggest that one explanation is a new shift from manufacturing toward services in advanced economies.

To bring new insights into the inequality–development literature, the current study applies penalized regression spline methods to top 1% income share data. The World Wealth and Income Database (formerly the World Top Incomes Database) provides unprecedentedly long inequality series that cover almost a century for many countries (Alvaredo et al., 2013b, 2016). During this period, some countries have faced not only urbanization but also more advanced stages of development. Due to data unavailability, the focus of the study is on "advanced" countries; however some "less-advanced" countries are also included in the total sample of 26 countries. The data are of high quality compared to many other inequality data. Moreover, Leigh (2007) and Roine and Waldenström (2015) provide evidence that these series reflect changes in other inequality indices over time. Thus, it is interesting to exploit top income shares in inequality–development studies, particularly when other alternatives for long series are not available.[4]

This study finds that the inequality–development association is U-shaped after a certain development level when inequality is measured in terms of the top 1% income share. In an additional investigation encompassing the years 1980–2009, the positive association (at the highest levels of economic development) is robust to including controls for urbanization and the service sector. Moreover, there are similarities in the overall shape of the inequality–development relationship when one compares the results of this paper to the pooled results in Frazer (2006), although the studies use different distributional measures.

The remainder of this study is organized in the following manner: Section 2 introduces the data used in the empirical analysis, and section 3 describes the estimation method. Section 4 provides empirical results including sensitivity analysis. Finally, section 5 presents the conclusions.

2. Data

2.1. Top 1% income shares

Many of the available Gini series have suffered from comparability problems, both in time and between countries, and the series have not covered long time intervals. Using tax and population statistics, it is possible to compose long and fairly consistent series on top income shares. Kuznets (1953) was the first to use this kind of data to produce top income share estimates, and Piketty (2001, 2003) generalized Kuznets's approach. Following Piketty, different researchers have constructed top income share series using similar methods.[5] According to Leigh (2007), the evolution of top income shares is similar to that of various other inequality indices over time. In addition, Roine and Waldenström (2015) conclude that top income shares are useful in describing inequality.

Top income data can be easily accessed using the World Wealth and Income Database by Alvaredo et al. (2016) (formerly the World Top Incomes

[4] To the best of the author's knowledge, there are no previous studies that exploit the new top income share series in this context.

[5] For more information on the methodology see, for example, Atkinson (2007). In addition, the advantages and limitations of the top income share series are discussed by Piketty and Saez (2006), Leigh (2007), and Roine and Waldenström (2015). Furthermore, Atkinson et al. (2011) provide a thorough overview of the top income literature.

Database by Alvaredo et al., 2013b).[6] The top 1% income shares in 26 countries from 1900 to 2010 are exploited, but the longitudinal data are not balanced (note that this is pre-tax income). Most of the data are from the English-speaking, Continental European, Southern European, and Nordic countries; however Japan, Singapore, and some "less-advanced" countries are also included.[7] The top 1% income share (*top1*) series are presented graphically in Appendix A. Table 1 provides summary statistics.

On the basis of the existing top income literature, an inverse U-shaped association between *top1* and economic development is not expected. For example, in the English-speaking countries, the evolution of the top 1% income shares resembles U over the twentieth century because there has been a significant increase since the 1980s; whereas the top 1% shares in Continental Europe and Japan have remained fairly stable during the past three decades. Further, Atkinson et al. (2011) and Roine and Waldenström (2015) discuss the problems of fitting the evolution of top income shares into the approach where the inequality–development relation is described by sectoral shifts. Other factors—also indicated by Kuznets (1955)—seem relevant, particularly taxation and the concentration of savings at the top.[8] Moreover, "superstar" theories and the possibility of changing norms are examples of suggested explanations for the recent increase in top incomes in many countries. For more discussion, see, for example, Piketty and Saez (2006) and Alvaredo et al. (2013a).

2.2. Economic development and sectoral variables

The level of economic development is measured in a traditional manner using GDP per capita. The GDP per capita data (1990 international GK$) are available annually until 2010 in the Maddison Project update (Bolt & van Zanden, 2013). Data from 1900 are used whenever available. In an additional analysis encompassing the years 1980–2009, the models include controls for

[6]The first book on these series, edited by Atkinson and Piketty (2007), contrasted the evidence from the Continental Europe and English-speaking countries. The second volume, also edited by Atkinson and Piketty, was published in 2010. The database builds on these volumes, and the project is ongoing.

[7]Argentina, China, Colombia, India, Indonesia, Mauritius, and South Africa.

[8]Roine et al. (2009) provide empirical evidence for the negative association between tax progressivity and top income shares. Moreover, Kanbur (2000) notes that inequality–development studies tend to minimize the role of policy.

Table 1: Descriptive statistics.

Annual data (1900–2010)	N	min	mean	max
top1	1609	2.7	10.6	28.0
ln(GDP p.c.)	1609	6.4	8.9	10.4
Data averaged over 5-year periods (1980–2009)[a]	N	min	mean	max
top1	129	3.0	8.8	20.5
ln(GDP p.c.)	129	7.2	9.5	10.3
urbanization	129	22.1	71.1	100.0
service sector	129	17.7	62.5	78.6

[a]The 5-year periods are defined as 1980–84, 1985–89, ..., and 2005–09.

two sectors, namely, urban and service sectors. It should be interesting to see whether the inclusion of sectoral variables affects the relationship between top-end inequality and economic development. Urbanization data describe the *population residing in urban areas (%)* (United Nations, 2012). These data are available every five years. The service sector is measured with *employment in service sector (% of total employment)* (World Bank, 2014a), and these data are available from 1980 onward. See Table 1 for descriptive statistics.

Although the investigated time span becomes considerably shorter with the two sectoral variables, this approach can be considered an extension to previous studies. For example, Frazer (2006) reports controlling for urbanization but does not provide detailed results on the inequality–urbanization relationship. Desbordes and Verardi (2012) do not include sectoral variables in their empirical models.[9]

3. Estimation method

Additive models provide a flexible framework for investigating the association between inequality and development.[10] This study follows the approach

[9]Kanbur and Zhuang (2013) is a recent example of focusing on the inequality–urbanization relationship in four Asian countries in the spirit of Kuznets (1955).

[10]Additive models are a special case of generalized additive models (GAMs). GAMs were introduced by Hastie and Tibshirani (1986, 1990). They present a GAM as a generalized linear model with a linear predictor that involves a sum of smooth functions of covariates. This study uses an identity link and assumes normality in errors, which leads to additive models.

presented in Wood (2006). The basic idea is that the model's predictor is a sum of linear and smooth functions of covariates:

$$\mathbb{E}(Y_i) = \boldsymbol{X}_i^* \boldsymbol{\theta} + f_1(x_{1i}) + f_2(x_{2i}) + ...$$

In the above presentation, Y_i is the response variable (here: *top1*), $\boldsymbol{X}_i^*$ is a row of the model matrix for any strictly parametric model components, $\boldsymbol{\theta}$ is the corresponding parameter vector, and the $f_\bullet$ are smooth functions of the covariates, $x_\bullet$.

The flexibility of these models comes at the cost of two problems. First, one needs to represent the smooth functions $f_\bullet$ in some manner. One way to represent these smooths is to use cubic regression splines, which is the approach adopted in this study. A cubic regression spline is a curve constructed from sections of cubic polynomials that are joined together so that the resulting curve is continuous up to the second derivative. The points at which sections are joined (and the end points) are the knots of the spline, and these locations must be chosen. The spline can be represented in terms of its values at the knots.[11] Second, the amount of smoothness that functions $f_\bullet$ will have needs to be chosen. Overfit is to be avoided and, thus, departure from smoothness is penalized. The appropriate degree of smoothness for $f_\bullet$ can be estimated from the data by, for example, maximum likelihood.

Illustration

Consider a model containing only one smooth function of one covariate: $y_i = f(x_i) + \epsilon_i$, where ϵ_i are i.i.d. $N(0, \sigma^2)$ random variables. To estimate function f here, f is represented so that the model becomes a linear model. This is possible by choosing a basis, defining the space of functions of which f (or a close approximation to it) is an element. In practice, one chooses basis functions, which are treated as known.

Assume that the function f has a representation $f(x) = \sum_{j=1}^{k} \beta_j b_j(x)$, where β_j are unknown parameters and $b_j(x)$ are known basis functions. Using a chosen basis for f implies that we have a linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where the model matrix $\boldsymbol{X}$ can be represented using basis functions such as those in the cubic regression spline basis. The departure from smoothness can be penalized with $\int f''(x)^2 dx$. The penalty $\int f''(x)^2 dx$ can be expressed as

[11] There are usually two extra conditions that specify that the second derivative of the curve should be zero at the two end knots.

$\boldsymbol{\beta}^T \boldsymbol{S}\boldsymbol{\beta}$, where $\boldsymbol{S}$ is a coefficient matrix that can be expressed in terms of the known basis functions.

Accordingly, the penalized regression spline fitting problem is to minimize $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta}$, with respect to $\boldsymbol{\beta}$. The problem of estimating the degree of smoothness is a problem of estimating the smoothing parameter λ.[12] The penalized least squares estimator of $\boldsymbol{\beta}$, given λ, is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{S})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. Thus, the expected value vector is estimated as $\widehat{\mathbf{E}(\boldsymbol{y})} = \hat{\boldsymbol{\mu}} = \boldsymbol{A}\boldsymbol{y}$, where $\boldsymbol{A} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{S})^{-1}\boldsymbol{X}^T$ is called an influence matrix.

This setting can be augmented to include several covariates and smooths. Given a basis, an additive model is simply a linear model with one or more associated penalties.

Practical notes

The size of basis dimension for each smooth is usually not critical in estimation, because it only sets an upper limit on the flexibility of a term. Smoothing parameters control the effective degrees of freedom (*edf*). Effective degrees of freedom are defined as trace($\boldsymbol{A}$), where $\boldsymbol{A}$ is the influence matrix. The effective degrees of freedom can be used to measure the flexibility of a model. It is also possible to divide the effective degrees of freedom into degrees of freedom for each smooth. For example, a simple linear term would have one degree of freedom, and *edf*=2.3 can be thought of as a function that is slightly more complex than a second-degree polynomial.

Confidence (credible) intervals for the model terms can be derived using Bayesian methods, and approximate *p*-values for model terms can be calculated. Models can be compared using information criteria such as the Akaike information criterion (AIC). When using the AIC for penalized models (models including smooth terms), the degrees of freedom are the effective degrees of freedom, not the number of parameters. Moreover, random effects can be included in these models. For further details, see Wood (2006).[13]

[12] In the estimation, one faces a bias–variance tradeoff: on the one hand, the bias should be small, but on the other hand, the fit should be smooth. One needs to compromise between the two extremes. $\lambda \longrightarrow \infty$ results in a straight line estimate for f, and $\lambda = 0$ leads to an unpenalized regression spline estimate.

[13] The results presented in this study are obtained using the R software package "mgcv" (version 1.7-21), which includes a function "gam." Basis construction for cubic regression splines is used (the knots are placed evenly through the range of covariate values by default). The maximum likelihood method is used in the selection of the smoothing

4. Results

In the baseline models, the estimation is implemented with annual data from 1900 to 2010. The results are also checked by studying different subsets of the sample and changing the data structure from annual to 5-year average data. Finally, in an additional analysis, urbanization and service sector variables are included in models with 5-year average data spanning the years from 1980 to 2009.

4.1. Baseline models

The baseline results are for annual data spanning 1900–2010. The models are of the form

$$top1_{it} = \alpha + f(ln(GDP\ p.c.)_{it}) + \delta_{decade} + u_i + \epsilon_{it},$$

where i refers to country and t to year, α is a constant, f is a smooth function that is described using a penalized cubic regression spline, δ_{decade} is a time dummy (one decade is the reference category), u_i is a country effect, and $\epsilon_{it} \sim N(0, \sigma^2)$ is the error term. The country effects can be omitted, fixed (i.e., dummy variables), or random ($u_i \sim N(0, \sigma_u^2)$). Different strategies in modeling country effects are reported because the literature does not follow a unified approach. Thus, the reader can also see when and how the chosen specification affects the results.

Details of the model without country effects are provided in column (1) of Table 2. Models (2) and (3) of Table 2 include country effects, and the table shows that including these effects improves the model fit. Figure 1 illustrates the smooth functions f in these three models. The fixed-effect (FE) and random-effect (RE) specifications give practically identical fits. In all three specifications, there is a possibility of a flat curve at lower levels of development (ln(GDP per capita) < 8, approximately). Further, after a certain level of development (ln(GDP per capita) > 8.5, approximately), all smooths show U shape (or J shape).[14]

parameters. The identifiability constraints (due to, for example, the model's additive constant term) are taken into account by default. The function "gam" also allows for simple random effects: it represents the conventional random effects in a GAM as penalized regression terms. More details can be found in Wood (2006) and the R project's web pages (`http://cran.r-project.org/`).

[14] Note: $\exp(8) \approx 2980$ and $\exp(8.5) \approx 4910$ (1990 international GK\$).

Table 2: Baseline models, using annual data (years 1900–2010): effective degrees of freedom for each smooth. Intercepts, country effects, and time effects[a] are not reported. For graphical illustration of smooth functions f, see Figure 1.

	dependent variable: $top1_t$ (N=1609) (1)	(2)	(3)
$f(ln(GDP\ p.c.)_t)$	[$edf \approx 9.2^{\mathrm{b}}$]*** See Fig. 1 (a)	[$edf \approx 10.4^{\mathrm{b}}$]*** See Fig. 1 (b)	[$edf \approx 10.4^{\mathrm{b}}$]*** See Fig. 1 (c)
country effects	no	fixed	random
AIC	7950	6642	6642

*** indicates significance at the 1% level.
The smooth terms' significance levels are based on approximate p-values.

[a] All models (1)–(3) include time effects. Time effects are dummy variables for different decades. However, all observations for 2000–2010 are considered in the "last" decade.
[b] The basis dimension of the smooth before imposing identifiability constraints is $k = 15$.

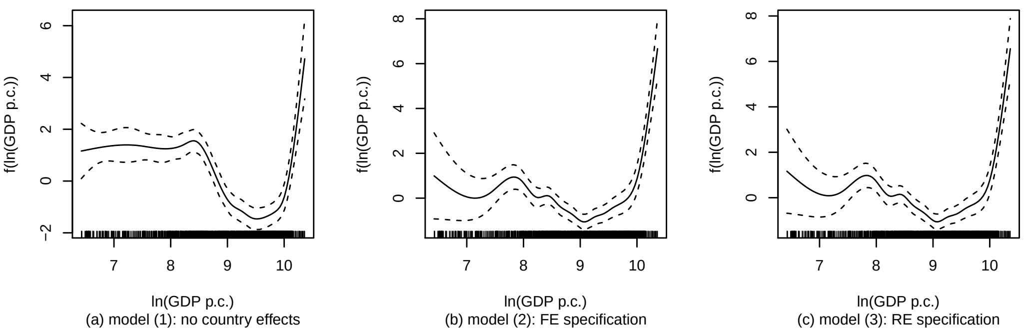

Figure 1: Illustration of the *top1*–development relation (annual data 1900–2010). See Table 2 for details. The solid line represents the smooth function $f(ln(GDP\ p.c.))$. The plots also show the 95% Bayesian credible intervals (dashed) and the covariate values as a rug plot along the horizontal axis.

The overall shape of $f(ln(GDP\ p.c.))$ resembles the shape that Frazer (2006) shows for the Gini–development relationship (except for the steep positive slope at the highest levels of development). This similarity supports the notion that top income shares reflect the same characteristics as the traditional Gini coefficients. Even the downward peak close to ln(GDP per capita) ≈ 9.5 in plot (a) of Figure 1 appears to be reasonable compared to Frazer's pooled models.[15]

4.2. Sensitivity of the baseline models' results

In the first check, the English-speaking, Nordic, Continental and Southern European, and "less-advanced" countries were studied separately.[16] More detailed information on the models with random country effects is reported in Table B.5 in Appendix B. The illustrations of the smooths $f(ln(GDP\ p.c.))$ in these specifications are provided in Figure 2. Plots in Figure 2 illustrate that the association is not uniform at lower levels of development (ln(GDP per capita) < 8.5, approximately). However, there seems to be a pattern that holds as countries reach a higher level of economic development: there is a negative relationship between *top1* and the level of development when $8.5 <$ ln(GDP per capita) < 9.5 (approximately). In general, the shape of the association between top-end inequality and development is fairly uniform when ln(GDP per capita) > 8.5. The results in Figure 2 are also in line with plot (c) of Figure 1, which illustrates the corresponding random-effect specification with the entire sample. Moreover, the main results of the fixed-effect specifications for separate groups accorded with those in Figure 2.[17]

The second check was concerned with the sensitivity of excluding groups of countries from the entire sample. The previously discovered U shape (or J shape) emerges again at development levels ln(GDP per capita) > 8.5 (approximately), and the downward peak of the U is located between $9 <$ ln(GDP per capita) < 10. More detailed information on the models with random country effects is provided in Figure B.5 in Appendix B. Further-

[15]Note: $\exp(9.5) \approx 13360$ (1990 international GK\$ in the current study).

[16]Singapore and Japan do not fit into these categories and were, thus, not included in these group-wise investigations.

[17]Only in the group of Continental and Southern European countries the curve may be flat at the highest levels of economic development. Detailed results on the fixed-effect specifications are not provided for the sake of brevity.

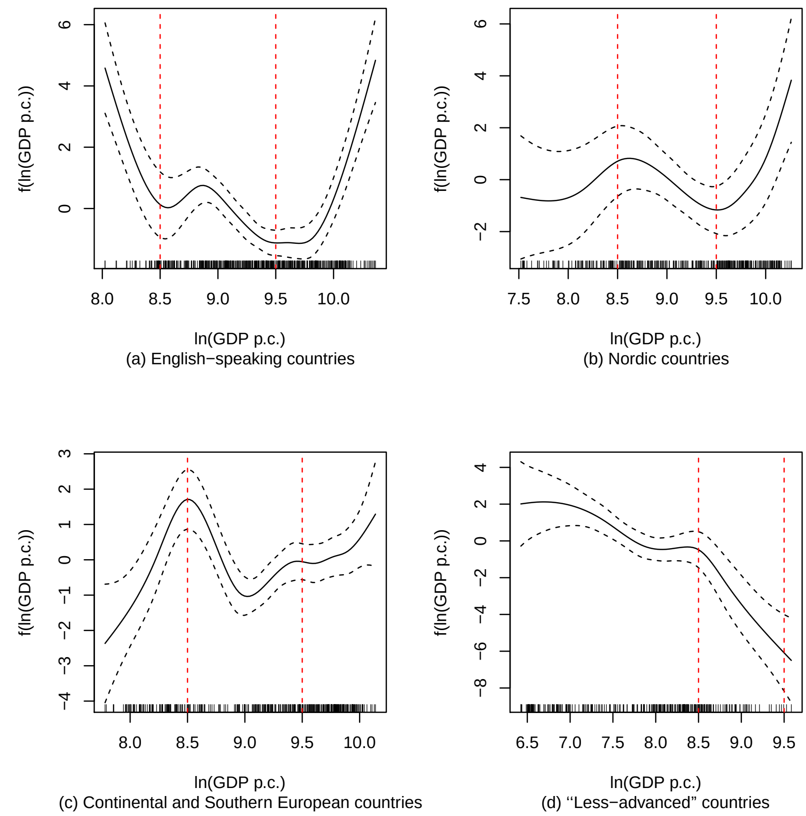

Figure 2: Illustration of the *top1*–development relation with four different subsets of the sample (annual data 1900–2010). The models include decade dummies and random country effects (Table B.5 in Appendix B provides details). The solid line represents the smooth function $f(ln(GDP\ p.c.))$. The plots also show the 95% Bayesian credible intervals (dashed) and the covariate values as a rug plot along the horizontal axis. Vertical dashed lines have been added to highlight the idea of a negative slope between $8.5 <$ ln(GDP per capita) < 9.5 (approximately).

more, when the corresponding fixed-effect specifications were studied, the results were similar to those reported in Figure B.5.[18]

Finally, the annual data results were checked against the corresponding results with the 5-year average data. The main findings with the 5-year data spanning 1900–2009 do not contradict the results presented in subsection 4.1. Appendix C provides graphical illustrations. Thus, the overall results do not seem to depend on the choice between annual and 5-year average data. The next investigations are conducted with 5-year averages, but the models are augmented with sectoral variables.

4.3. Additional analysis: controlling for two sectors

This subsection provides an additional analysis where models include controls for urban and service sectors. The analysis is implemented using 5-year averages, where the periods are 1980–1984, 1985–1989, ..., and 2005–2009.[19] The studied specifications are as given below:

$$\begin{aligned} top1_{it} &= \alpha + f_1(ln(GDP\ p.c.)_{it}) + f_2(urbanization_{it}) \\ &+ f_3(service\ sector_{it}) + \delta_{decade} + u_i + \epsilon_{it}, \end{aligned}$$

where i refers to country and t to 5-year period, α is a constant, smooth functions f_j $(j = 1, 2, 3)$ are approximated using penalized cubic regression splines, δ_{decade} is a fixed time effect (one decade is the reference category), u_i is a country effect, and $\epsilon_{it} \sim N(0, \sigma^2)$ is the error term; the values of the top 1% share, ln(GDP per capita) and sectoral variables are now period averages. As before, the country effects can be omitted, fixed, or random depending on the specification. Initially, all smooths f_j were allowed to enter in a flexible form, but a linear term was suggested for the service sector variable in some models. The models in question were then re-estimated with this linearity restriction.

Table 3 provides details on models with two sectors. Models (2) and (3) have linear terms for the service sector, and the coefficients are provided

[18]The detailed results on the fixed-effect specifications are not reported for the sake of brevity. In addition, the effect of excluding Japan and Singapore from the sample was tested because these two countries do not fit into the discussed categorization. The main results that relate to "medium" and "high" levels of development are not sensitive to including or excluding these countries.

[19]Taking period averages should reduce potential short-run disturbances. Moreover, the urbanization variable is available every five years.

Table 3: Models with two sectors, using 5-year average data (years 1980–2009): effective degrees of freedom for each smooth $f_\bullet$ and coefficients for linear terms. Intercepts, country effects, and time effects[a] are not reported. The smooths with $edf > 1$ are illustrated in Figure 3.

	dependent variable: $top1_t$ (N=129)		
	(1)	(2)	(3)
$f_1(ln(GDP\ p.c.)_t)$	[$edf \approx 4.7^{b}$]*** See Fig. 3 (a)	[$edf \approx 5.3^{b}$]*** See Fig. 3 (b)	[$edf \approx 5.4^{b}$]*** See Fig. 3 (c)
$f_2(urbanization_t)$	[$edf \approx 5.8^{b}$]*** See Fig. 3 (d)	[$edf \approx 3.6^{b}$]* See Fig. 3 (e)	[$edf \approx 4.1^{b}$]** See Fig. 3 (f)
$f_3(service\ sector_t)$	[$edf \approx 2.9^{b}$]*** See Fig. 3 (g)	[linear[b]] 0.096**	[linear[b]] 0.120***
country effects	no	fixed	random
AIC	542	361	371

***, **, * indicate significance at the 1, 5, and 10% levels, respectively.
The p-values for parametric terms are calculated using the Bayesian estimated covariance matrix of the parameter estimators; only the significance levels are reported. The smooth terms' significance levels are based on approximate p-values.
[a]All models (1)–(3) include time effects. Time effects are dummy variables for different decades.
[b]The basis dimension of the smooth before imposing identifiability constraints is $k = 10$.

in the table. Figure 3 provides plots of the other smooth functions. The results on sectoral variables are fairly uniform, irrespective of the country-effect specification. The urbanization smooth resembles an inverted-U curve (particularly in plots (e) and (f) of Figure 3). The association between the top 1% share and employment in services is positive, which leads to speculation regarding whether this illustrates a new structural shift.

Let us then focus on the GDP per capita variable. In plots (a) and (c) of Figure 3, the model without country effects and the model with random country effects show very similar shapes for the smooth $f(ln(GDP\ p.c.))$, and the overall shape does not contradict previously reported results.[20] In contrast, the fixed-effect specification in plot (b) does not confirm the U-shaped relationship at "medium-to-high" levels of development.[21] However,

[20]Moreover, Frazer (2006) controls for urbanization in the sensitivity checks of his pooled model and finds that the overall shape of the Gini–development relationship holds.

[21]This conclusion regarding the smooth $f(ln(GDP\ p.c.))$ does not change if the sectoral variables are excluded from the model with fixed country effects (when period 1980–2009 is studied).

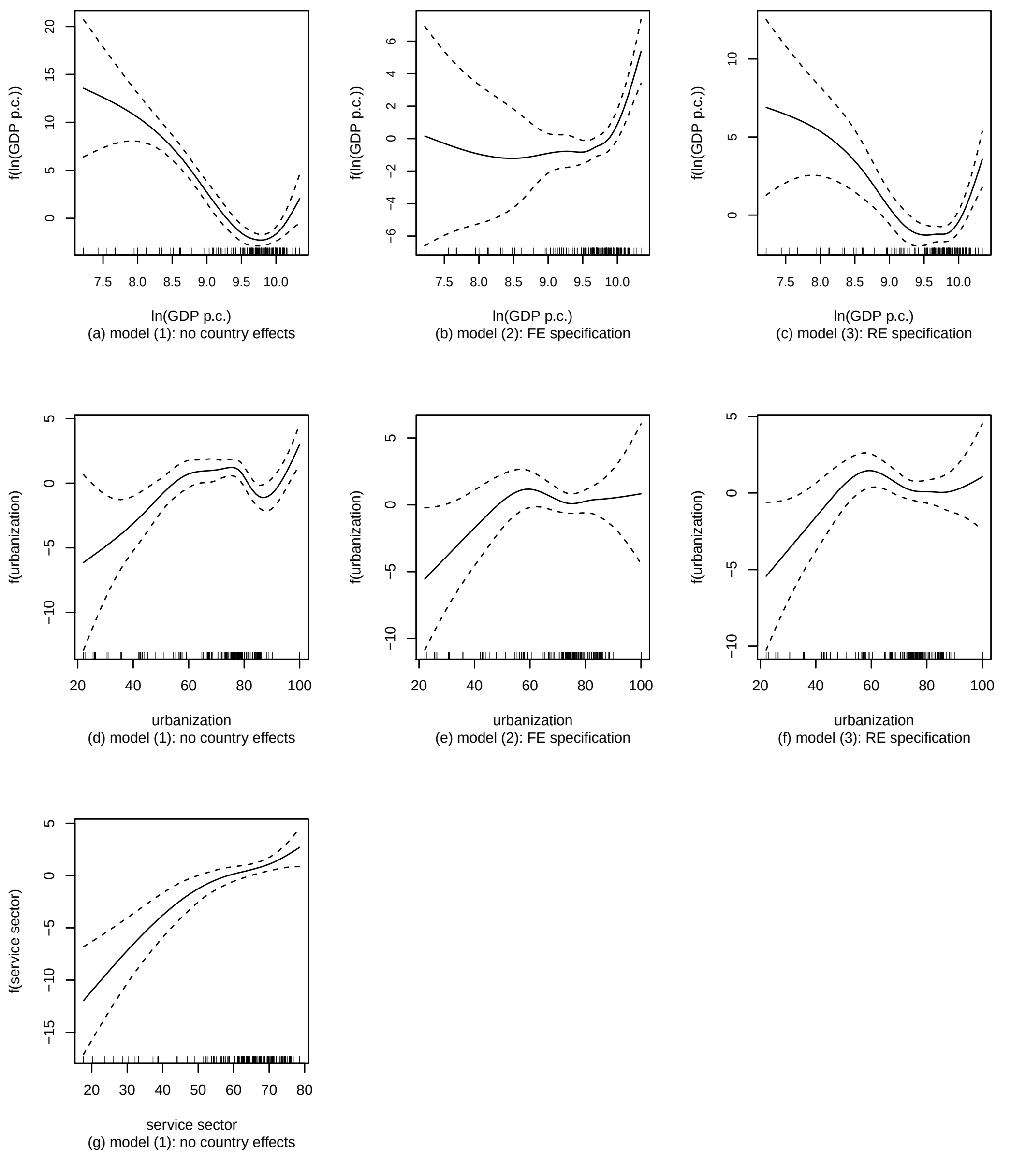

Figure 3: Illustrations of the smooths, using 5-year average data (years 1980–2009). See Table 3 for the details of the models. The solid line represents the smooth function $f_\bullet$. The plots also show the 95% Bayesian credible intervals (dashed) and covariate values as a rug plot along the horizontal axis.

the positive relationship at the highest levels of GDP per capita is discovered in all three specifications, and the "turning point" is located close to ln(GDP per capita) $\approx$ 9.5. Thus, the discovered positive association holds at the highest levels of development when two sectors are controlled for.

These results were also checked against leaving country groups out of the sample, one group at a time. The categorization was the same as that in the previous subsection (and in the lower panel of Table B.5). The random-effect specifications were intuitive when compared to the model (3) of Table 3. In comparison, the fixed-effect specifications were slightly more sensitive to the exclusion of country groups, but also these findings were reasonable when compared to the whole-sample results of model (2) in Table 3.[22] For brevity, the details of these checks are not reported.

Finally, an alternative measure for the service sector was tested. Data on *services, etc., value added (% of GDP)* (World Bank, 2014b) begin from the 1960s for some countries, but Swiss data are not available. Results related to ln(GDP per capita) and urbanization did not change. The alternative service sector measure correlated positively with *top1*, but it was not statistically significant at the 10% level in specifications with country effects. However, these results were not in conflict with the models of Table 3. Thus, the details of these checks are not reported.

5. Discussion

A vast number of empirical studies have explored the relationship between inequality and development, but the results have been mixed. The current paper addresses the issue by applying flexible methods to new data. The results of the current study are based on an unbalanced longitudinal data from 26 countries over the years 1900–2010. Various specifications in this paper suggest a negative association between the top 1% income share and

[22]Main findings with the FE specifications: When the Continental and Southern European countries were excluded from the sample, GDP per capita variable was not statistically significantly related to the top 1% share at the 10% level; both sectoral variables correlated positively with the top 1% share. In comparison, when the "less-advanced" countries were excluded, the sectoral variables were not significantly related to the top 1% share at the 10% level, but—as expected—there was a statistically significant, positive relationship between per capita GDP and the top income share. Further, excluding either the English-speaking or the Nordic countries from the sample barely affected the main conclusions.

ln(GDP per capita) after a certain point in the development process. Furthermore, the current study finds that this relationship turns positive at even higher levels of economic development. Thus, the data suggest a reversal of the Kuznets curve after a certain development level is reached. However, the current sample includes only some "less-advanced" countries, and more research is needed when new data become available.

In an additional analysis encompassing the period 1980–2009, this study assumes a broad interpretation of Kuznets's idea of sectoral shifts. The analysis is descriptive, but the results favor that something more than sectoral shifts are needed to explain changes in top-end inequality in the course of economic development. Specifically, the discovered positive association between the top 1% share and economic development (at the highest levels of development) holds when measures for urbanization and service sector are included. This accords with the existing literature on top incomes, which has highlighted other explanations for the evolution of top income shares.

Appendix A. Top 1% income share series

Table A.4: Top 1% income share series (years 1900–2010). For better comparability, series excluding capital gains have been selected whenever possible. Figure A.4 plots the series. For more information, see Atkinson and Piketty (2007, 2010) and Alvaredo et al. (2016).

Country	***N***	**Source**
Argentina	39	Alvaredo et al. (2013b)
Australia	90	Alvaredo et al. (2013b)
Canada	91	Alvaredo et al. (2013b)[b]
China	18	Alvaredo et al. (2013b)
Colombia	18	Alvaredo et al. (2013b)
Denmark	95	Alvaredo et al. (2013b)
Finland	90	Alvaredo et al. (2013b)[c]
France	96	Alvaredo et al. (2013b)[d]
Germany	47	Alvaredo et al. (2013b)
India	71	Alvaredo et al. (2013b)
Indonesia	28	Alvaredo et al. (2013b)
Ireland	37	Alvaredo et al. (2013b)
Italy	34	Alvaredo et al. (2013b)
Japan	110	Alvaredo et al. (2013b)
Mauritius	52[a]	Alvaredo et al. (2013b)
Netherlands	55	Alvaredo et al. (2013b)
New Zealand	83	Alvaredo et al. (2013b)
Norway	69	Alvaredo et al. (2013b)
Portugal	24	Alvaredo et al. (2013b)
Singapore	59[a]	Alvaredo et al. (2013b)
South Africa	62[a]	Alvaredo et al. (2013b)
Spain	30	Alvaredo et al. (2013b)
Sweden	79	Alvaredo et al. (2013b)
Switzerland	74	Alvaredo et al. (2013b)[e]
United Kingdom	60	Alvaredo et al. (2013b)
United States	98	Alvaredo et al. (2013b)
	total: 1609	

[a] There would be more top 1% income share observations, but GDP per capita data are not available: Mauritius (+4), Singapore (+3), and South Africa (+9).

[b] Two partially overlapping series are available. Here; series up to 1981 is based on tax data, and series from 1982 is based on Longitudinal Administrative Database.

[c] Two partially overlapping series are available. Here; series up to 1989 is based on tax data, and the series from 1990 is based on the Income Distribution Survey.

[d] In the original source, the figure for 1905 is averaged for 1900–1910.

[e] For all years except 1933, the estimates relate to income averaged over the year shown and the following year. Thus, repeated value for two consecutive years is used in this study.

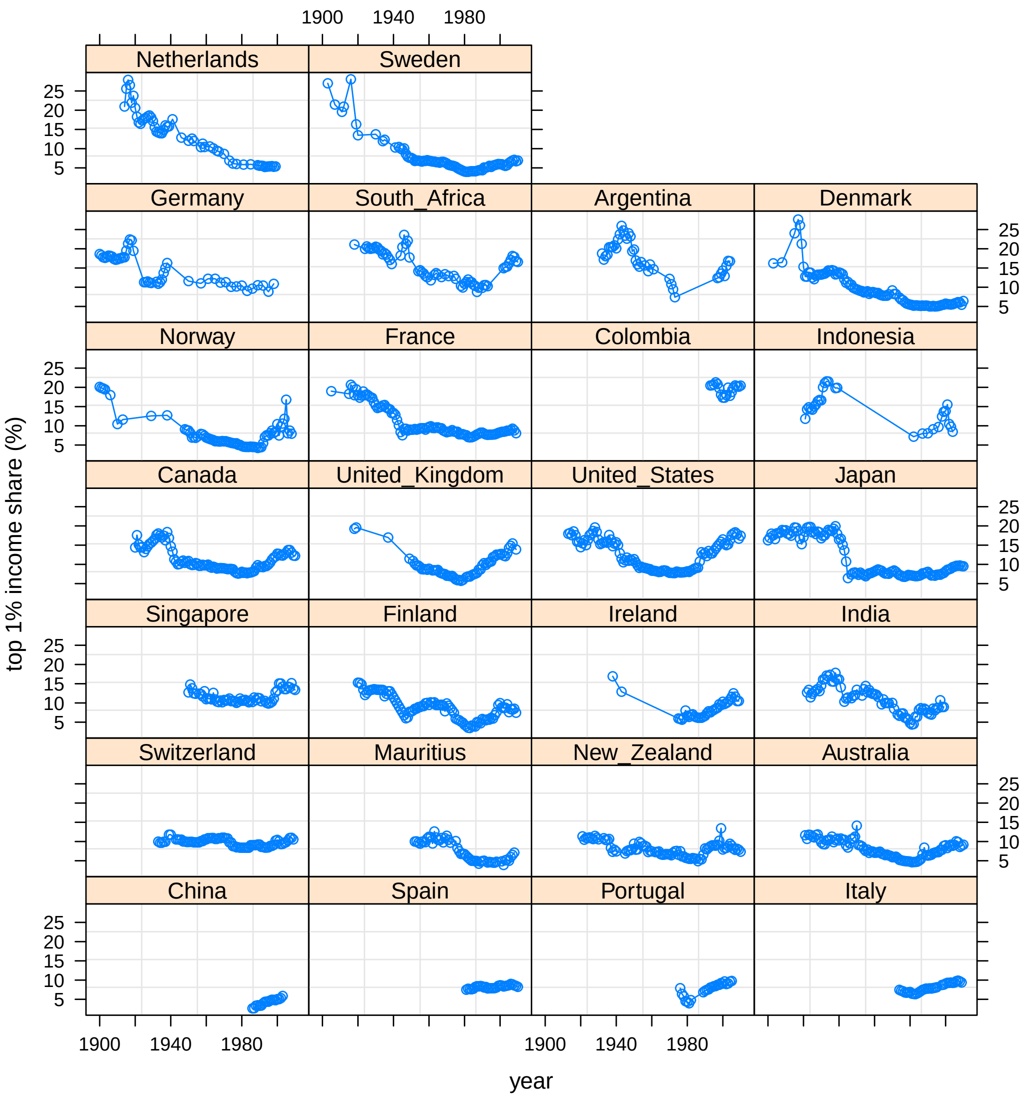

Figure A.4: Top 1% income share series for each country (years 1900–2010). See Table A.4 for details. Data source: Alvaredo et al. (2013b).

Appendix B. Model details: subsets of the sample

Table B.5: Subsets of the sample. Results of models with fixed time effects[a] and random country effects, using annual data (years 1900–2010): effective degrees of freedom for each smooth.

$top1_{it} = \alpha + f(ln(GDP\ p.c.)_{it}) + \delta_{decade} + u_i + \epsilon_{it}$	N	smooth f
In Figure 2:		
(a) English-speaking [b]	459	[$edf \approx 7.4^{\mathrm{f}}$]***
(b) Nordic [c]	333	[$edf \approx 5.9^{\mathrm{f}}$]***
(c) Continental and Southern Europe [d]	360	[$edf \approx 6.9^{\mathrm{f}}$]***
(d) "Less-advanced" [e]	288	[$edf \approx 5.4^{\mathrm{f}}$]***
In Figure B.5:		
(a) Without English-speaking [b]	1150	[$edf \approx 10.0^{\mathrm{g}}$]***
(b) Without Nordic [c]	1276	[$edf \approx 10.0^{\mathrm{g}}$]***
(c) Without Continental/Southern Europe [d]	1249	[$edf \approx 9.8^{\mathrm{g}}$]***
(d) Without "less-advanced" [e]	1321	[$edf \approx 9.5^{\mathrm{g}}$]***

*** indicates significance at the 1% level.
The smooth terms' significance levels are based on approximate p-values.

[a] Time effects are dummy variables for different decades. However, all observations for 2000–2010 are considered in the "last" decade. One decade is the reference category.
[b] Australia, Canada, Ireland, New Zealand, the United Kingdom, and the United States.
[c] Denmark, Finland, Norway, and Sweden.
[d] France, Germany, Italy, the Netherlands, Portugal, Spain, and Switzerland.
[e] Argentina, China, Colombia, India, Indonesia, Mauritius, and South Africa.
[f] The basis dimension of the smooth before imposing identifiability constraints is $k = 10$.
[g] The basis dimension of the smooth before imposing identifiability constraints is $k = 15$.

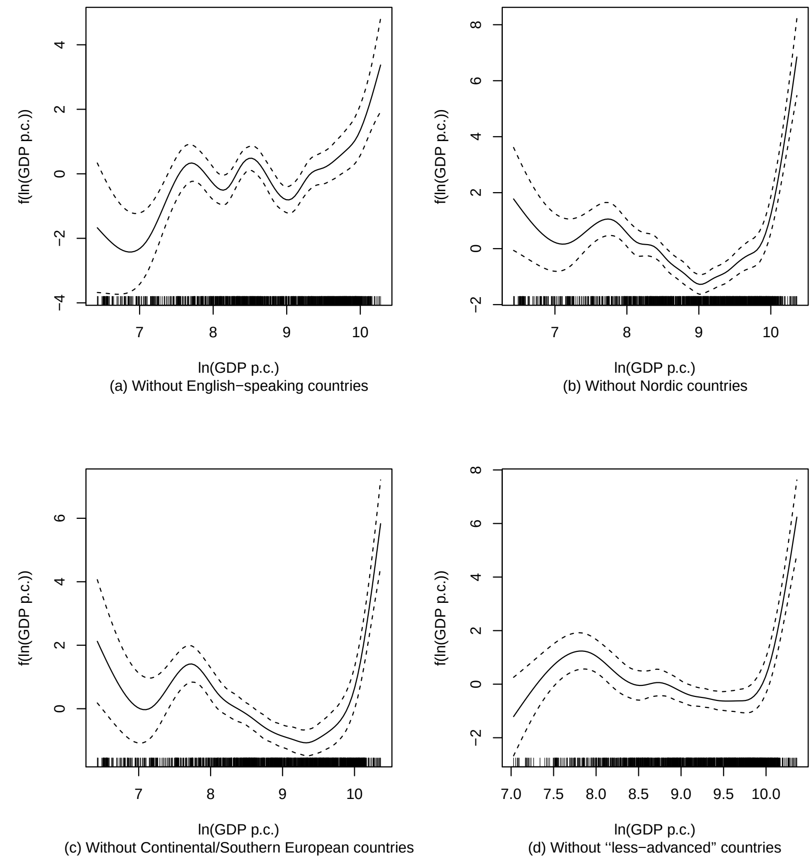

Figure B.5: The effect of leaving country groups out of the sample (annual data 1900–2010). The models include decade dummies and random country effects. See Table B.5 for model details. The solid line represents the smooth function $f(ln(GDP\ p.c.))$. The plots also show the 95% Bayesian credible intervals (dashed) and covariate values as a rug plot along the horizontal axis. The shapes of these smooths can be compared to plot (c) of Figure 1, which illustrates the corresponding random-effect specification with the entire sample.

Appendix C. 5-year average data: results using the long series

The baseline models with the 5-year average data (discussed at the end of subsection 4.2) are of the form $top1_{it} = \alpha + f(ln(GDP\ p.c.)_{it}) + \delta_{decade} + u_i + \epsilon_{it}$, where i refers to country and t to 5-year period,[23] α is a constant, f is a smooth function that is described using a penalized cubic regression spline, δ_{decade} is a fixed time effect (one decade is the reference category), u_i is a country effect (omitted, fixed, or random), and ϵ_{it} is the conventional error term; the values for top 1% share and ln(GDP per capita) refer to period averages. Figure C.6 below describes the smooths f.[24] The obtained shapes of $f(ln(GDP\ p.c.))$ are close to the corresponding ones in Figure 1. Thus, changing the modeling strategy from annual to 5-year average data does not influence the overall shapes of the corresponding smooths.

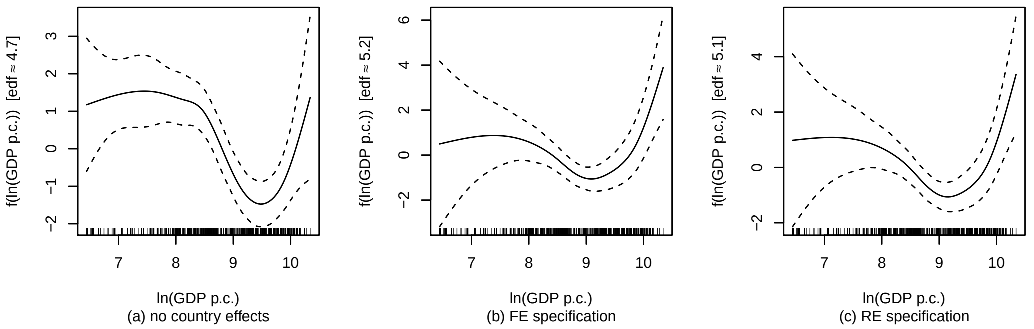

Figure C.6: Illustration of the *top1*–development relation, using 5-year average data (years 1900–2009, here N=376). The solid line represents the smooth function $f(ln(GDP\ p.c.))$. The figure also shows the 95% Bayesian credible intervals (dashed), and covariate values as a rug plot along the horizontal axis. Plot (a) represents a model without country effects, plot (b) illustrates a model with country-specific fixed effects, and plot (c) represents a model with country-specific random effects. All models include decade dummies.

[23]These periods are 1900–04, 1905–09, ..., and 2005–09.

[24]The basis dimension of the smooth before imposing identifiability constraints is $k = 10$.

References

Aghion, P., Bolton, P., 1997. A Theory of Trickle-Down Growth and Development. Review of Economic Studies 64(2), 151–172.

Ahluwalia, M.S., 1976. Inequality, Poverty and Development. Journal of Development Economics 3(4), 307–342.

Alvaredo, F., Atkinson, A.B., Piketty, T., Saez, E., 2013a. The Top 1 Percent in International and Historical Perspective. Journal of Economic Perspectives 27(3), 3–20.

Alvaredo, F., Atkinson, A.B., Piketty, T., Saez, E., 2013b. The World Top Incomes Database. Data downloaded from website: `http://g-mond.parisschoolofeconomics.eu/topincomes` (September 12, 2013).

Alvaredo, F., Atkinson, A.B., Piketty, T., Saez, E., Zucman, G., 2016. The World Wealth and Income Database. Website: `http://www.wid.world`.

Anand, S., Kanbur, S.M.R., 1993. Inequality and development: A critique. Journal of Development Economics 41(1), 19–43.

Atkinson, A.B., 2007. Measuring Top Incomes: Methodological Issues, in: Atkinson, A.B., Piketty, T. (Eds.), Top Incomes over the Twentieth Century: A Contrast between Continental European and English-Speaking Countries. Oxford University Press, Oxford. pp. 18–42.

Atkinson, A.B., Brandolini, A., 2001. Promise and Pitfalls in the Use of "Secondary" Data-Sets: Income Inequality in OECD Countries as a Case Study. Journal of Economic Literature 39(3), 771–799.

Atkinson, A.B., Piketty, T. (Eds.), 2007. Top Incomes over the Twentieth Century: A Contrast between Continental European and English-Speaking Countries. Oxford University Press, Oxford.

Atkinson, A.B., Piketty, T. (Eds.), 2010. Top Incomes: A Global Perspective. Oxford University Press, Oxford.

Atkinson, A.B., Piketty, T., Saez, E., 2011. Top incomes in the Long Run of History. Journal of Economic Literature 49(1), 3–71.

Barro, R.J., 2000. Inequality and Growth in a Panel of Countries. Journal of Economic Growth 5(1), 5–32.

Bolt, J., van Zanden, J.L., 2013. The First Update of the Maddison Project; Re-Estimating Growth Before 1820. Maddison Project Working Paper 4. Data downloaded from website: `http://www.ggdc.net/maddison/maddison-project/data/mpd_2013-01.xlsx` (September 12, 2013).

Dahan, M., Tsiddon, D., 1998. Demographic Transition, Income Distribution, and Economic Growth. Journal of Economic Growth 3(1), 29–52.

Deininger, K., Squire, L., 1996. A New Data Set Measuring Income Inequality. World Bank Economic Review 10(3), 565–591.

Deininger, K., Squire, L., 1998. New ways of looking at old issues: Inequality and growth. Journal of Development Economics 57(2), 259–287.

Desbordes, R., Verardi, V., 2012. Refitting the Kuznets curve. Economics Letters 116(2), 258–261.

Fields, G.S., 2001. Distribution and Development: A New Look at the Developing World. The MIT Press, Cambridge, Massachusetts.

Frazer, G., 2006. Inequality and Development Across and Within Countries. World Development 34(9), 1459–1481.

Galor, O., Tsiddon, D., 1996. Income Distribution and Growth: the Kuznets Hypothesis Revisited. Economica 63(250), S103–S117.

Greenwood, J., Jovanovic, B., 1990. Financial Development, Growth and the Distribution of Income. Journal of Political Economy 98(5), 1076–1107.

Hastie, T., Tibshirani, R., 1986. Generalized additive models (with discussion). Statistical Science 1(3), 297–318.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall/CRC, New York.

Huang, H.–C.R., 2004. A flexible nonlinear inference to the Kuznets hypothesis. Economics Letters 84(2), 289–296.

Huang, H.–C.R., Lin, S.–C., 2007. Semiparametric Bayesian inference of the Kuznets hypothesis. Journal of Development Economics 83(2), 491–505.

Kanbur, R., 2000. Income Distribution and Development, in: Atkinson, A.B., Bourguignon F. (Eds.), Handbook of Income Distribution Vol. 1. North-Holland, Amsterdam. pp. 791–841.

Kanbur, R., Zhuang, J., 2013. Urbanization and Inequality in Asia. Asian Development Review 30(1), 131–147.

Kuznets, S., 1953. Shares of Upper Income Groups in Income and Saving. NBER Publication No. 55, New York.

Kuznets, S., 1955. Economic Growth and Income Inequality. American Economic Review 45(1), 1–28.

Leigh, A., 2007. How Closely Do Top Income Shares Track Other Measures of Inequality? Economic Journal 117(524), F589–F603.

Lin, S.–C., Huang, H.–C.R., Weng, H.-W., 2006. A semi-parametric partially linear investigation of the Kuznets' hypothesis. Journal of Comparative Economics 34(3), 634–647.

List, J.A., Gallet, C.A., 1999. The Kuznets Curve: What Happens After the Inverted-U? Review of Development Economics 3(2), 200–206.

Piketty, T., 2001. Les Hauts revenus en France au 20e siècle: inégalités et redistribution, 1901–1998. B. Grasset, Paris.

Piketty, T., 2003. Income Inequality in France 1901–1998. Journal of Political Economy, 111(5), 1004–1042.

Piketty, T., Saez, E., 2006. The Evolution of Top Incomes: A Historical and International Perspective. American Economic Review 96(2), 200–205.

Robinson, S., 1976. A Note on the U Hypothesis Relating Income Inequality and Economic Development. American Economic Review 66(3), 437–440.

Roine, J., Vlachos, J., Waldenström, D., 2009. The long-run determinants of inequality: What can we learn from top income data? Journal of Public Economics 93(7–8), 974–988.

Roine, J., Waldenström, D., 2015. Long-Run Trends in the Distribution of Income and Wealth, in: Atkinson, A.B., Bourguignon, F. (Eds.), Handbook of Income Distribution Vol. 2A. North-Holland, Amsterdam. pp. 469–592.

United Nations, 2012. Department of Economic and Social Affairs, Population Division. World Urbanization Prospects: The 2011 Revision (POP/DB/WUP/Rev.2011/1/F2). Data downloaded from website: `http://esa.un.org/unpd/wup/CD-ROM/WUP2011-F02-Proportion_Urban.xls` (November 28, 2013).

Wood, S.N., 2006. Generalized Additive Models: An Introduction with R. Chapman & Hall/CRC, Boca Raton FL.

World Bank, 2014a. World Development Indicators, Employment in services (% of total employment). Data downloaded from website: `http://data.worldbank.org/indicator/SL.SRV.EMPL.ZS/` (February 25, 2014).

World Bank, 2014b. World Development Indicators, Services, etc., value added (% of GDP). Data downloaded from website: `http://data.worldbank.org/indicator/NV.SRV.TETC.ZS/` (February 25, 2014).

Zhou, X., Li, K.–W., 2011. Inequality and development: Evidence from semiparametric estimation with panel data. Economics Letters 113(3), 203–207.